Universidade Federal do Paraná (UFPR)
Bacharelado em Informática Biomédica

Aprendizagem Bayesiana

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

David Menotti
www.inf.ufpr.br/menotti/ci171-182

Aprendizagem Bayesiana

Agenda

- Introdução
- Teorema de Bayes
- Classificador Naïve Bayes

Aprendizagem Bayesiana

- O pensamento Bayesiano fornece uma abordagem probabilística para a aprendizagem.
- Está baseado na suposição de que as quantidades de interesse são reguladas por distribuições de probabilidades.
- Quantificar o custo/benefício entre diferentes decisões de classificação usando probabilidades e custos associados à classificação.
- Teorema de Bayes
  – Mostra como alterar as probabilidades *a priori* tendo em conta novas evidências de forma a obter probabilidades *a posteriori*.

Aprendizagem Bayesiana

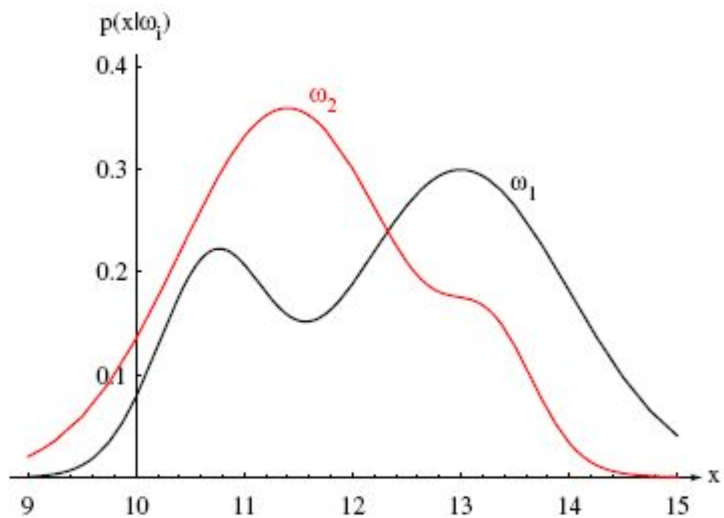
Terminologia

- Classes ω_i (variável aleatória)
- Probabilidades *a priori* $P(\omega_i)$
  - Conhecimento *a priori* que se tem sobre o problema, ou seja, conhecimento *a priori* sobre a aparição de exemplos das classes do problema.
- Função de Densidade Probabilidade $P(x)$
  - Frequência com a qual encontramos uma determinada característica
  - Evidências

Aprendizagem Bayesiana

Terminologia

- Densidade de Probabilidade Condicional
  - $P(x|\omega_j)$ (*Likelihood*) - Verossimilhança
  - Frequência com que encontramos uma determinada característica *x* dado que a mesma pertence a classe ω_j



Densidade de duas classes em que x representa uma característica qualquer

Aprendizagem Bayesiana

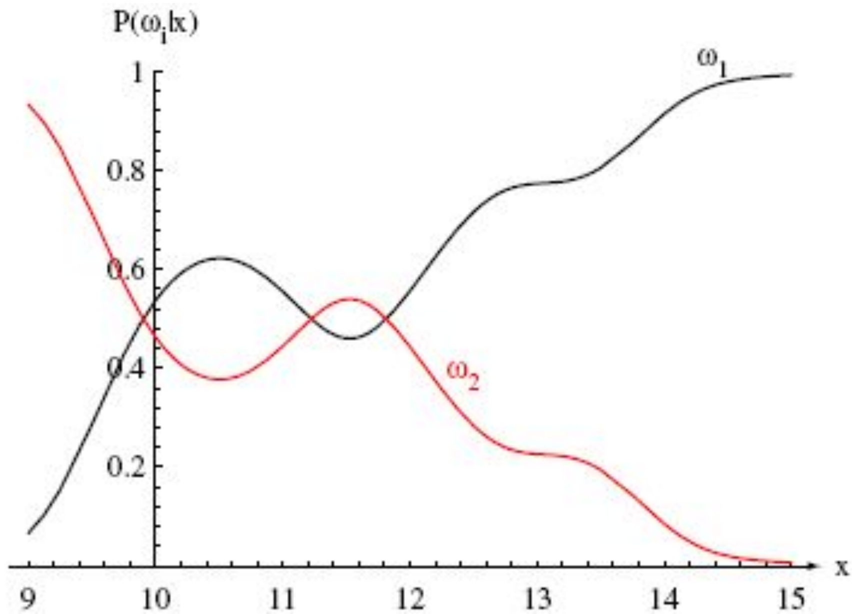
Terminologia

- Probabilidade *a posteriori*
  - $P(\omega_j|x)$ - Probabilidade que o padrão pertença a classe ω_j data a característica x

- Regra de decisão
  - ω_1, se $P(\omega_1) > P(\omega_2)$
  - ω_2, caso contrário

Aprendizagem Bayesiana

Tomando decisão usando Bayes



Probabilidades *a posteriori* calculadas usando $P(\omega_1) = 2/3$ e $P(\omega_2) = 1/3$

Nesse caso, para um valor de $x = 14$,

a probabilidade do padrão pertencer a ω_1 é de 0,08, enquanto que

a probabilidade do padrão pertencer a ω_2 é de 0,92.

Para cada x, as probabilidades *a posteriori* somam 1.

Aprendizagem Bayesiana

Teorema de Bayes

- Basicamente o teorema de Bayes mostra como rever as crenças sempre que novas evidências são coletadas.
- Ou seja, atualizar a probabilidade *a posteriori* utilizando para isso a probabilidade *a priori*, as verossimilhanças e as evidências

$$P(A|B) = \frac{P(A) \times P(B|A)}{P(B)}$$

- $P(A|B)$ é a probabilidade a posteriori
- $P(A)$ é a probabilidade a priori
- $P(B|A)$ são as verossimilhanças $(likelihood)$
- $P(B)|$ são as evidências, dado por $\sum P(A_i) \times P(B|A_i)$

Aprendizagem Bayesiana

Exemplo

- Um médico sabe que a meningite causa torcicolo em 50% dos casos. Porém, o médico sabe que a meningite atinge 1/50.000 e também que a probabilidade de se ter torcicolo é de 1/20.
- Usando Bayes para saber a probabilidade de uma pessoa ter meningite dado que ela está com torcicolo

Temos então

- $P(\,T \mid M\,) = 0{,}5$ $\qquad P(\,M\,) = 1 / 50.000$ $\qquad P(\,T\,) = 1 / 20$

$$P(M|T) = \frac{P(M)\times P(T|M)}{P(T)} = \frac{\frac{1}{50000}\times 0.5}{1/20}$$

$$P(M|T) = 0,0002 = \frac{2}{10000} = 0,02\%$$

Aprendizagem Bayesiana

Exercício

- Considere o sistema de classificação de peixes. Para essa época do ano, sabe-se que a probabilidade de pescar salmão é maior que pescar robalo, P(salmao) = 0,82 e P(robabo) = 0,18.
- Suponha que a única característica que você pode contar é a intensidade do peixe ou seja, se ele é claro ou escuro. Sabe-se que 49,5% dos salmões tem intensidade clara e que 85% dos robalos tem intensidade clara.
- Calcule a probabilidade de ser salmão dado que o peixe pescado tem intensidade clara.

$$P(S|C) = \frac{P(S)\times P(C|S)}{P(C)} = \frac{0{,}82\times 0{,}495}{0{,}82\times 0{,}495+0{,}18\times 0{,}85} = 0,726$$

Classificador Naïve Bayes

- Um dos algoritmos de aprendizagem mais práticos e utilizados na literatura.

- Denominado Naive (ingênuo) por assumir que os atributos são condicionalmente independentes, ou seja, a informação de um evento não é informativa sobre nenhum outro.

- Apesar dessa premissa, o classificador reporta bom desempenho em diversas tarefas de classificação onde há dependência.

- Aplicações bem sucedidas:
  – Diagnóstico médico
  – Classificação de documentos textuais

Classificador Naïve Bayes

- Aplica-se a tarefas de aprendizagem onde cada instância x é descrita por uma conjunção de valores de atributos em que a função alvo, $f(x)$, pode assumir qualquer valor de um conjunto V

- Um conjunto de exemplos de treinamento da função alvo é fornecido. E então uma nova instância é apresentada, descrita pela tupla de valores de atributos $< a_1, a_2, \ldots, a_n >$

- A tarefa é predizer o valor alvo (ou classificação) para esta nova instância.

Classificador Naïve Bayes

- O classificador é baseado na suposição de que os valores dos atributos são **condicionalmente independentes** dados o valor alvo.
- Se usarmos Bayes para múltiplas evidências, temos

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1, E_2, \ldots, E_n|H) \times P(H)}{P(E_1, E_2, \ldots, E_n)}$$

- Considerando a hipótese de independência, podemos reescrever o teorema de Bayes da seguinte forma:

$$P(H|E_1, E_2, \ldots, E_n) = \frac{P(E_1|H) \times, P(E_2|H) \times, \ldots, P(E_n|H) \times P(H)}{P(E_1, E_2, \ldots, E_n)}$$

O denominador pode ser ignorado por se tratar de um termo comum.

Classificador Naïve Bayes

Exemplo - Considere o seguinte problema:

	outlook	temperature	humidity	windy	play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

Classificador Naïve Bayes

Construindo o Modelo (NB)

- O primeiro passo consiste em construir o modelo de probabilidades condicionais Naïve Bayes (NB) - **14 dias**

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								

- A tabela acima contém a frequência de diferentes evidências.
- Por exemplo, existem duas instâncias mostrando (***Outlook*** = `Sunny`) para (***Play*** = `Yes`)

Classificador Naïve Bayes

Computando probabilidades

- Após definir todas as frequências é necessário calcular todas as probabilidades condicionais e as probabilidades *a priori*.

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- Por exemplo
  - *P*(***Outlook*** = `Sunny` | ***`Play`*** = `Yes`) = 2/9
  - *P*(***`Play`*** = `Yes`) = 9/14

Classificador Naïve Bayes

Predição

- De posse do modelo, podemos usá-lo para predizer um evento "***Play***"' com base em um conjunto qualquer de evidências.
- Por exemplo: `[Sunny,Cool,High,True,?]`

P(*Yes* | *E*) = (*P*(***Outlook*** = `Sunny` | *Yes*) x
P(***Temp*** = `Cool` | *Yes*) x
P(***Humidity*** = `High` | *Yes*) x
P(***Windy*** = `True` | *Yes*) x
P(*Yes*)) / *P*(*E*)

P(*E*) pode ser ignorada por se tratar de um denominador comum quando queremos comparar as duas classes. Deste modo, temos

$$P(Yes|E) = \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}$$

Classificador Naïve Bayes

Predição

- Calculando a predição para as duas classes
  - Para `Yes` temos $\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14} = 0,0053$
  - Para `No` temos $\frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14} = 0,0206$

- Convertendo esses valores para probabilidade através da normalização, temos
  - *P*(`Yes` | *E*) = 0,0053 / (0,0053+0.0206) = 0,205 = 20,5%
  - *P*(`No` | *E*) = 0,0206 / (0,0053+0.0206) = 0,795 = 79,5%

Classificador Naïve Bayes

Técnica de Suavização

- Em alguns casos, a frequência pode ser zero, como por exemplo

  P(***Outlook*** = `Overcast` | ***`Play`*** = *`No`*) = 0 / 5

- Isso cria um problema para calcular *P*(*`No`*), a qual será sempre zero quando esta evidência for utilizada.
  – Toda instância que contiver esta evidência terá probabilidade nula.

Classificador Naïve Bayes

Técnica de Suavização

- A técnica de suavização mais utilizada é a estimação de Laplace

$$P'(H|E) = \frac{n_c + \mu p}{n + \mu}$$

- n_c é o número de hipóteses existentes para a classe (Ex: Zero para ***Outlook*** = `Overcast` e *Play* = *No*)
- n número de exemplos totais para o treinamento
- Considerado que as evidências são igualmente distribuídas, tempo $p = \frac{1}{3}$ (`Sunny` , `Overcast` , `Rainy`)
- μ número de exemplos virtuais (distribuição uniforme?)

Classificador Naïve Bayes

Técnica de Suavização

- Reestimando os valores usando Laplace, teríamos
  - *P*(***Outlook*** = `Sunny` | ***`Play`*** = *`No`*) = (3 + 3x1/3) / (5+3) = 4 / 8
  - *P*(***Outlook*** = `Overcast` | ***`Play`*** = *`No`*) = (0 + 3x1/3) / (5+3) = 1 / 8
  - *P*(***Outlook*** = `Rainy` | ***`Play`*** = *`No`*) = (2 + 3x1/3) / (5+3) = 3 / 8
- Desta forma, todos os valores foram redistribuídos mantendo uma proporção similar

Classificador Naïve Bayes

Calculando as probabilidades para atributos contínuos

Existem duas maneiras

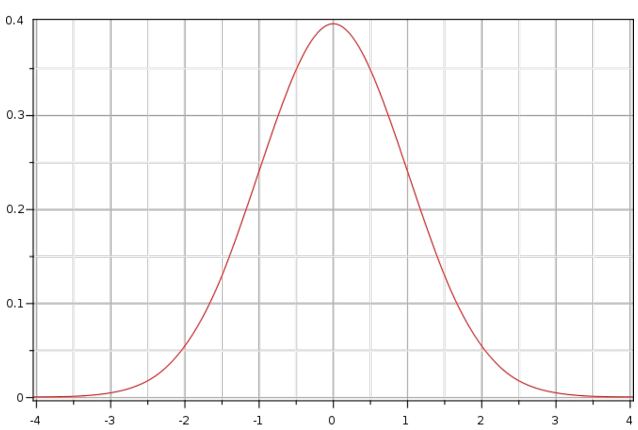
- Discretizar os atributos contínuos em algumas categorias, usando conhecimento sobre o domínio / aplicação.
  - Por exemplo
    - Temperatura acima de 80F pode ser considerada alta.
    - Idade acima de 18 anos, entre 18 e 65 anos, acima dos 65 anos
    - Salário anual: 50k, 100k, 200k, 500k, 1M
    - Gordura total: <100, [100,200], >200

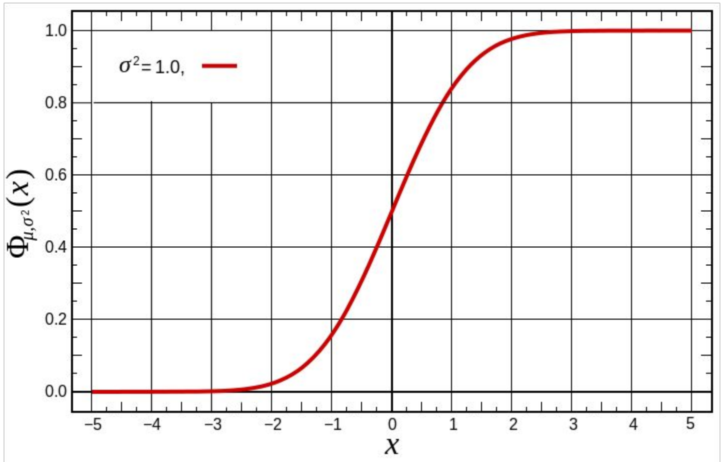
Classificador Naïve Bayes

Calculando as probabilidades para atributos contínuos

- Outra forma consiste em usar uma função de densidade de probabilidade (PDF) e desta forma preservar os valores contínuos.
  - Nesse caso assumimos que as variáveis contínuas seguem uma **distribuição normal**
  - Com isso em mente, podemos calcular a **média** e **desvio padrão** de cada variável usando a base de aprendizagem.
  - De posse da média e desvio padrão, basta aplicar a fórmula da normal para estimar a probabilidade

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

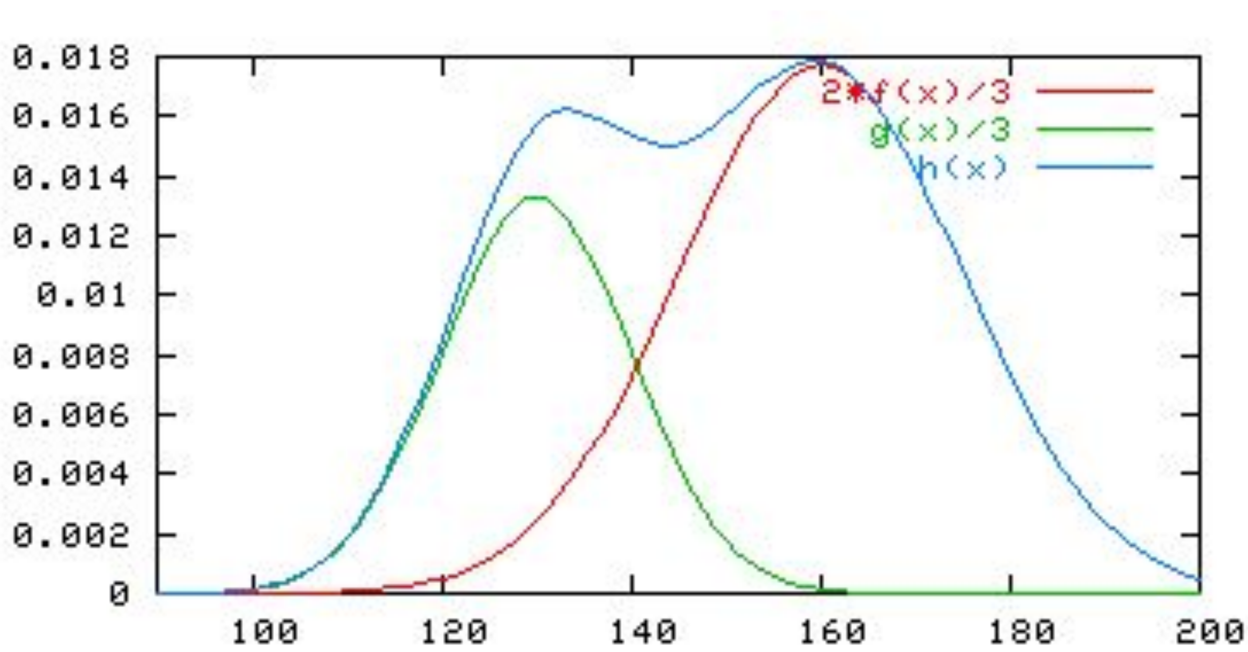




Classificador Naïve Bayes

Calculando as probabilidades para atributos contínuos

- Mistura de Gaussianas (GMM)
  - É possível ajustar várias Gaussianas aos dados



Classificador Naïve Bayes

Exemplo

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal distribution

		Humidity	*Mean*	*StDev*
Play Golf	yes	86 96 80 65 70 80 70 90 75	79.1	10.2
	no	85 90 70 95 91	86.2	9.7

(2*pi)^(-0.5) = 0,4

~0,4

$e^{-0,125}$ =0,882

$$P(\text{humidity} = 74 \mid \text{play} = \text{yes}) = \frac{1}{\sqrt{2\pi}(10.2)} e^{-\frac{(74-79.1)^2}{2(10.2)^2}} = 0.0344$$

$e^{-0,79}$ =0,453

$$P(\text{humidity} = 74 \mid \text{play} = \text{no}) = \frac{1}{\sqrt{2\pi}(9.7)} e^{-\frac{(74-86.2)^2}{2(9.7)^2}} = 0.0187$$

P(C|h)=P(C)P(h|C)

P(yes)=9/14
P(no) =5/14

Classificador Naïve Bayes

Exemplo

Attribute	Class yes (0.63)	no (0.38)
outlook		
sunny	3.0	4.0
overcast	5.0	1.0
rainy	4.0	3.0
[total]	12.0	8.0
temperature		
mean	72.9697	74.8364
std. dev.	5.2304	7.384
weight sum	9	5
precision	1.9091	1.9091
humidity		
mean	78.8395	86.1111
std. dev.	9.8023	9.2424
weight sum	9	5
precision	3.4444	3.4444
windy		
TRUE	4.0	4.0
FALSE	7.0	3.0
[total]	11.0	7.0

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Calcular a probabilidade para

E = [***Outlook*** = Rainy, ***Temp*** = 65, ***Humid*** = 70, ***Wind*** = True]

P(Yes | E) = 4 / 12 x 0,023 x 0,0271 x 4/11 x 0,63 $\Rightarrow$ 0,75 = 75%

P(No | E) = 3 / 8 x 0,022 x 0,0094 x 4/7 x 0,38 $\Rightarrow$ 0,25 = 25%

Classificador Naïve Bayes

Exemplo de Aplicação

- Classificação de texto, identificação de autoria, etc
  - Considere por exemplo que seja necessário construir um classificador que discrimine entre documentos relevantes e não relevantes

  - Base de treinamento deve conter documentos das duas classes

  - Quais características devem ser usadas?
    - *Bag of Words*: Conjunto pré-definido de palavras
    - Frequência de aparição dessas palavras pode ser utilizada como características

Referências

- Luiz E. S Oliviera, **Aprendizagem Bayesiana**, DInf / UFPR, 2017.